Am J Hum Genet 25:471–492, 1973

Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characters

JOSEPH FELSENSTEIN[1]

When we try to reconstruct the evolutionary tree of a group of organisms by examining a series of characters, we are not applying strict logical deduction but are making a guess in the presence of uncertainty. It is therefore appropriate to think of the problem in terms of statistical inference. This approach was first suggested by Edwards and Cavalli-Sforza [1–4]. The data collected by systematists and by students of molecular evolution are mostly for discrete characters, such as the presence or absence of a morphological structure or the amino acid sequence of a protein. But much data are also collected for quantitative characters, such as gene frequencies and measurements on morphological traits. In this paper, I will confine my attention to quantitative characters. This is the case originally considered by Edwards and Cavalli-Sforza. They proposed that the estimation of evolutionary trees be carried out by the method of maximum likelihood. However, they found troublesome singularities in what they believed to be the likelihood surface [3, 4]. They were forced to fall back on ad hoc approaches which did not have an explicit statistical justification (their "method of minimum evolution" and "additive tree model"; see also [5]). Malyutov et al. [6] have described another ad hoc approach.

In this paper, I will use the basic model proposed by Edwards and Cavalli-Sforza. I will show that if we are less ambitious than they were, and redefine the problem so as not to attempt to estimate as many quantities, we can construct a likelihood function which does not have any such singularities. It is then possible to construct computer programs which obtain maximum-likelihood estimates of the evolutionary tree when the data are in the form of quantitative measurements.

TREES

Before the model is described in detail, it may be helpful to consider what is meant by the term "evolutionary tree." If we knew nothing of the amount and quality of real data available, we might wish to know the entire evolutionary history

Received August 9, 1972; revised December 18, 1972.

A portion of this work was submitted as part of a Ph.D. thesis to the Department of Zoology, University of Chicago.

This study was supported by NIH fellowships 5T01 GM-00090 and 1-F2-GM-36,536-01, and task agreement no. 5 of contract AT(45-1) 2225 from the U.S. Atomic Energy Commission.

[1] Department of Genetics, University of Washington, Seattle, Washington 98195.

© 1973 by the American Society of Human Genetics. All rights reserved.

of a group of organisms, including the exact pedigree of their ancestors and the phenotype of every individual which has ever existed. Lack of sufficient relevant information immediately forces us to abandon this attempt. With no hope of discovering pedigrees, we must usually confine ourselves to an attempt to discover the pattern of speciation events (or their equivalent). With no hope of discovering individual phenotypes, we are restricted to statements about such quantities as population means, and only at certain specified moments (such as the times of the branchings). We are thus usually restricted to estimating the topological form of the tree, the times of the forks, and the mean population phenotypes at these nodes.

Cavalli-Sforza and Edwards [3] encountered singularities when they attempted to estimate all of these quantities at once. In doing computer iterations searching for the maximum-likelihood tree, they found that an infinite increase in their likelihood function could be achieved simply by taking any internal segment of a tree and shortening it to length zero. They were trying to estimate the times of branching, t_1 and t_2, at the ends of each segment, as well as the mean phenotypes, x and y, in those populations. The overall likelihood of a tree was computed as the product of probabilities corresponding to each segment of the tree. The term corresponding to a particular segment was of the form $f(y|x, t_1, t_2)dy$. The function $f\ dy$ is the probability that a population starting at phenotype x at time t_1 will change to a phenotype in the interval $(y, y + dy)$ between times t_1 and t_2. But if, in a tree, $t_1 = t_2$ and $x = y$, this probability is not an infinitesimal quantity but one. It is *certain* that in the time interval from t_1 to t_1, the population will "move" from x to x. Thus by setting a segment length to zero, we have achieved an infinite increase in "likelihood" of the tree, defined in this particular way. Furthermore, this "likelihood" increases continuously as the length of the segment in time is shortened and as y is made closer to x. Thus an iterative procedure for obtaining a maximum-likelihood tree will simply shorten one segment of the initial tree and finally "blow up" when the segment length reaches zero. Edwards [4] has since concluded that the function which they were calculating was not the likelihood. This is at least intuitively reasonable, since one would hope that a likelihood surface would not have a singularity unless the data give infinitely more support to one hypothesis than to any other.

One way of avoiding this problem is to make no attempt to estimate the phenotypes at the forks of the tree. We shall see that doing this, estimating only the topological form of the tree and the times of branching, leads directly to a likelihood formula without singularities. In this formula, there is almost no difference between the likelihood of a tree with a very short segment and a similar tree in which this segment has zero length.

THE MODEL

To keep derivations manageable, I will state an oversimplified model and will discuss its extension to more realistic cases later in this paper. The model is that used by Edwards and Cavalli-Sforza [1, 2]. We consider p characters evolving independently. As time passes, each character follows a Brownian motion, with a

mean displacement of zero and a variance in displacement of σ^2 per unit time. This means that after t units of time have elapsed, a phenotype can be considered to have changed by an amount drawn from a normal distribution with mean zero and variance $\sigma^2 t$. At the same time, the populations occasionally split. Immediately after the time of splitting, the two daughter populations have the same phenotypes at each character, but from that moment on the phenotypes in the two populations change independently by continuation of the Brownian motion process in each population. Our data consists of the phenotypes at the branch tips of a tree. The expression for the likelihood of a tree will be the probability density of this observed data given that evolution has taken place according to that particular tree. We will need a formula for this probability density.

As an example, consider the simple evolutionary tree shown in figure 1. The tip

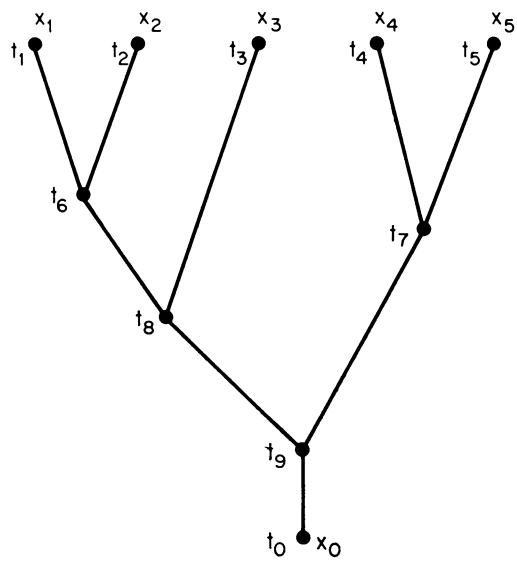

FIG. 1.—Tree used as example in text

populations are numbered 1–5, and the forks are numbered 6–9. The t_i represent the times of the nodes. The values x_1 through x_5 are the observed phenotypes of a character. Consider x_1 and x_3, as well as the unknown phenotypes at the forks below these populations, x_6, x_8, and x_9. Assume for the moment that the initial value x_0 of the character is known. Obviously,

$$x_1 = (x_1 - x_6) + (x_6 - x_8) + (x_8 - x_9) + (x_9 - x_0) + x_0, \qquad (1)$$

and

$$x_3 = (x_3 - x_8) + (x_8 - x_9) + (x_9 - x_0) + x_0. \qquad (2)$$

If we knew only the time, t_i, and the initial phenotype, x_0, what is the distribution of x_1 if it is generated by the Brownian motion process? Each of the successive

displacements $x_1 - x_6$, $x_6 - x_8$, $x_8 - x_9$, and $x_9 - x_0$ are drawn from normal distributions. Therefore their sum, x_1, is normally distributed with mean x_0. Its variance is the variance of the sum in equation (1). Since the component parts are independent, the variance of the sum is the sum of the variances:

$$\begin{aligned} \text{var}\ (x_1) &= \text{var}(x_1 - x_6) + \text{var}(x_6 - x_8) + \text{var}(x_8 - x_9) + \text{var}(x_9 - x_0) \\ &= \sigma^2(t_1 - t_6) + \sigma^2(t_6 - t_8) + \sigma^2(t_8 - t_9) + \sigma^2(t_9 - t_0) \\ &= \sigma^2(t_1 - t_0). \end{aligned} \tag{3}$$

Thus each of the phenotypes x_1, x_2, x_3, x_4, and x_5 is normally distributed with mean x_0 and variances $\sigma^2(t_1 - t_0)$, . . . , $\sigma^2(t_5 - t_0)$. Since each value is the sum of normally distributed variates, the set of values $(x_1, \ldots, x_5)$ follows a multivariate normal distribution. To completely characterize such a distribution, we need only know the means, variances, and covariances. We know the first two, so we want to obtain the covariances for all pairs of tip populations. Consider the covariance of x_1 and x_3. To calculate it, we must obtain the sum of all pairwise covariances between terms in equation (1) and terms in equation (2). All of these covariances are zero except when the same term appears in both equations. Then

$$\begin{aligned} \text{cov}(x_1, x_3) &= \text{cov}(x_8 - x_9, x_8 - x_9) + \text{cov}(x_9 - x_0, x_9 - x_0) \\ &= \text{var}(x_8 - x_9) + \text{var}(x_9 - x_0) \\ &= \sigma^2(t_8 - t_0). \end{aligned} \tag{4}$$

That all other terms are zero follows from the independence of the change in a character during different time periods.

Thus the distribution of $(x_1, \ldots, x_5)$ is multivariate normal, with means being x_0 and the covariance between populations i and j being $\sigma^2(t_k - t_0)$, where t_k is the time of the latest common ancestor of populations i and j. The covariance of two populations is thus proportional to the length of time they shared a common ancestor, and the variance of the distribution of a population's phenotype is proportional to the time it has been evolving. We can write the probability density of the vector of values $\mathbf{x} = (x_1, \ldots, x_n)$ as

$$f(\mathbf{x}|\mathbf{T},\mathbf{x}_0,\sigma^2) = \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{T}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{x}_0)(\sigma^2\mathbf{T})^{-1}(\mathbf{x} - \mathbf{x}_0)' \right] dx_1 dx_2 \ldots dx_n, \tag{5}$$

where $\mathbf{x}_0$ is the vector $(x_0, x_0, \ldots, x_0)$, n is the number of tip populations, and $\mathbf{T}$ is the $n \times n$ matrix whose (i, j) element t_{ij} is the time of joint evolution of populations i and j.

Actually, we have data on p characters, not just on one. Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$ be the vector of phenotypes of character i in populations 1 through n. Let $x_0^{(i)}$ be the initial phenotype of character i at the root of the tree, and let $\mathbf{x}_0^{(i)}$ be the corresponding vector $(x_0^{(i)}, \ldots, x_0^{(i)})$. Since evolution in different characters is assumed to be independent, the probability density of the whole set of data, given

$\mathbf{T}$, the $x_0^{(i)}$, and σ^2, is given by the product of p expressions, each like the right-hand side of equation (5), giving

$$f = \frac{1}{(2\pi)^{np/2}|\sigma^2\mathbf{T}|^{p/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{p} (\mathbf{x}^{(i)} - \mathbf{x}_0^{(i)})(\sigma^2\mathbf{T})^{-1}(\mathbf{x}^{(i)} - \mathbf{x}_0^{(i)})' \right] dx_1 dx_2 \ldots dx_n. \tag{6}$$

Considered "a priori," holding $\mathbf{T}$, the $x_0^{(i)}$, and σ^2 constant and varying the phenotypes of the characters in the tip populations, equation (6) gives the probability density for each possible set of data. Considering equation (6) "a posteriori," holding the observed $x_j^{(i)}$ fixed and letting $\mathbf{T}$, σ^2, and the $x_0^{(i)}$ vary, it gives the likelihood of each combination of these quantities. It is this which forms the basis of maximum-likelihood estimation of evolutionary trees, as specified by $\mathbf{T}$.

The reader may be forgiven a certain skepticism at this point. If equation (6) gives the probability (actually, the probability density) of obtaining the observed data, why does it not contain any terms giving the probability that a random pattern of branching would give the tree $\mathbf{T}$? The logical status of such terms is at least questionable. While they seem to provide prior probabilities for the trees $\mathbf{T}$, any model of branching contains parameters, such as a rate of splitting, λ, which must themselves be estimated by maximum likelihood. Thus we do not have a true prior distribution but are still engaged in maximum-likelihood estimation. Under these circumstances, the use of such a pseudo-prior is questionable.

When such terms are inserted in a straightforward way, the resulting "likelihood" surface has singularities. This suggests that more careful consideration is needed before the branching process can be taken into account adequately. A model of generation of the tree by random branching is also unrealistic in at least three major respects. It fails to take into account the continual risk of extinction. It assumes that we are observing all surviving populations, whereas usually our data are for a sample of only a few populations out of many. Furthermore, these few populations may be a very biased sample of those which survived. For all of these reasons, I do not feel that it is appropriate to insert terms for the branching process into equation (6). Cavalli-Sforza and Edwards [3] came to the same conclusion, and I see no reason to dispute their judgment on this point.

In addition to $\mathbf{T}$, we are also estimating σ^2 and the $x_0^{(i)}$. I will attempt to give the $x_0^{(i)}$ a decent burial later in this paper. As for σ^2, it is completely confounded with $\mathbf{T}$ in most cases. If we have values of $\mathbf{T}$ and σ^2 which maximize the likelihood, we can double σ^2 and halve the t_{ij}. Since the product, $\sigma^2\mathbf{T}$, remains unchanged, and since σ^2 and $\mathbf{T}$ enter into equation (6) only in their product, the likelihood must necessarily also remain unchanged.

Unless we have prior knowledge of σ^2 or of some of the t_{ij}, we are actually estimating $\sigma^2\mathbf{T}$, not $\mathbf{T}$. It is often convenient to set $\sigma^2 = 1$ and to think of the resulting values of $\mathbf{T}$ as times measured in units of $1/\sigma^2$.

PRUNING THE TREE

Equation (6) is unusable on at least two grounds. The $x^{(i)}$ have been assumed to be independent, while the characters we actually observe are rarely independent. In addition, the likelihood expression is so cumbersome as to be nearly useless. It involves inverting a matrix each time a likelihood is calculated. Searching for the maximum-likelihood value of **T** would be a very slow process, even with a large computer. In this section, I will outline a faster method of computing the likelihood. I will show later in the paper that this method can be easily extended to deal with correlated characters.

Consider again the tree in figure 1. Pick two branch tips which are both immediate descendants of the same fork, for example, populations 1 and 2 (we could also have chosen 4 and 5). Both are immediate descendants of fork 6. Let $v_1 = t_1 - t_6$ and $v_2 = t_2 - t_6$ be the lengths of the segments leading to populations 1 and 2. Now we replace the two phenotype values for each character in populations 1 and 2 by two new variables, $u_1^{(i)}$ and $x_6^{(i)}$, defined by

$$u_1^{(i)} = x_1^{(i)} - x_2^{(i)}$$

and (7)

$$x_6^{(i)} = \left(\frac{v_2}{v_1 + v_2}\right) x_1^{(i)} + \left(\frac{v_1}{v_1 + v_2}\right) x_2^{(i)}.$$

We want to know the distribution of the new sets of phenotypes $\mathbf{x}^{(i)} = (u_1^{(i)}, x_6^{(i)}, x_3^{(i)}, x_4^{(i)}, x_5^{(i)})$. Since the $x_3^{(i)}$, $x_4^{(i)}$, and $x_5^{(i)}$ are unchanged, and since the $u_1^{(i)}$ and $x_6^{(i)}$ are linear combinations of normally distributed quantities, the distribution must be multivariate normal. To characterize the distribution, we need only know the expectations and the covariances of the new population phenotypes. The expectations and pairwise covariances of the $x_3^{(i)}$, $x_4^{(i)}$, and $x_5^{(i)}$ are unchanged. For the expectations of the $u_1^{(i)}$ and the $x_6^{(i)}$ we have

$$E(u_1^{(i)}) = E(x_1^{(i)}) - E(x_2^{(i)}) = x_0^{(i)} - x_0^{(i)} = 0$$

and

$$E(x_6^{(i)}) = \left(\frac{v_2}{v_1 + v_2}\right) E(x_1^{(i)}) + \left(\frac{v_1}{v_1 + v_2}\right) E(x_2^{(i)}) \qquad (8)$$

$$= \frac{v_2 x_0^{(i)} + v_1 x_0^{(i)}}{v_1 + v_2} = x_0^{(i)}.$$

For their variances, we have

$$\begin{aligned} \mathrm{var}(u_1^{(i)}) &= \mathrm{var}(x_1^{(i)}) + \mathrm{var}(x_2^{(i)}) - 2\mathrm{cov}(x_1^{(i)}, x_2^{(i)}) \\ &= \sigma^2(t_1 - t_0) + \sigma^2(t_2 - t_0) - 2\sigma^2(t_6 - t_0) \\ &= \sigma^2(t_1 + t_2 - 2t_6) \\ &= \sigma^2(v_1 + v_2), \end{aligned}$$

and

$$\begin{aligned}\operatorname{var}(x_6^{(i)}) &= \frac{v_2^2}{(v_1+v_2)^2}\operatorname{var}(x_1^{(i)}) + \frac{v_1^2}{(v_1+v_2)^2}\operatorname{var}(x_2^{(i)}) \\ &\quad - 2\frac{v_1v_2}{(v_1+v_2)^2}\operatorname{cov}(x_1^{(i)},x_2^{(i)}) \\ &= \sigma^2\left[\frac{v_2^2(t_1-t_0)+v_1^2(t_2-t_0)-2v_1v_2(t_6-t_0)}{(v_1+v_2)^2}\right] \qquad (9) \\ &= \sigma^2\left[t_6+\frac{v_2^2(t_1-t_6)+v_1^2(t_2-t_6)}{(v_1+v_2)^2}-t_0\right] \\ &= \sigma^2\left[t_6+\frac{v_1v_2}{(v_1+v_2)}-t_0\right].\end{aligned}$$

For the covariance of $u_1^{(i)}$ and $x_6^{(i)}$,

$$\begin{aligned}\operatorname{cov}(u_1^{(i)}, x_6^{(i)}) &= \left(\frac{v_2}{v_1+v_2}\right)\operatorname{cov}(x_1^{(i)}, x_1^{(i)}) - \left(\frac{v_2}{v_1+v_2}\right)\operatorname{cov}(x_1^{(i)}, x_2^{(i)}) \\ &\quad + \left(\frac{v_1}{v_1+v_2}\right)\operatorname{cov}(x_1^{(i)}, x_2^{(i)}) \\ &\quad - \left(\frac{v_1}{v_1+v_2}\right)\operatorname{cov}(x_2^{(i)}, x_2^{(i)}) \\ &= \sigma^2\left[\left(\frac{v_2}{v_1+v_2}\right)(t_1-t_0-t_6+t_0)\right. \qquad (10) \\ &\quad \left. - \left(\frac{v_1}{v_1+v_2}\right)(t_2-t_0-t_6+t_0)\right] \\ &= \sigma^2\left[\frac{v_2v_1-v_1v_2}{(v_1+v_2)}\right] = 0.\end{aligned}$$

There remain only the covariances of $u_1^{(i)}$ and $x_6^{(i)}$ with the other variables $x_3^{(i)}$, $x_4^{(i)}$, and $x_5^{(i)}$. These are immediate:

$$\begin{aligned}\operatorname{cov}(x_6^{(i)}, x_3^{(i)}) &= \left(\frac{v_2}{v_1+v_2}\right)\operatorname{cov}(x_1^{(i)}, x_3^{(i)}) \\ &\quad + \left(\frac{v_1}{v_1+v_2}\right)\operatorname{cov}(x_2^{(i)}, x_3^{(i)}) \\ &= \operatorname{cov}(x_1^{(i)},x_3^{(i)}) = \operatorname{cov}(x_2^{(i)}, x_3^{(i)}) \qquad (11)\end{aligned}$$

and

$$\operatorname{cov}(u_1^{(i)}, x_3^{(i)}) = \operatorname{cov}(x_1^{(i)}, x_3^{(i)}) - \operatorname{cov}(x_2^{(i)}, x_3^{(i)}) = 0,$$

the other covariances being analogous.

The $u_1^{(i)}$ have zero covariance with all other population phenotypes, and each is normally distributed with mean zero and variance $\sigma^2(v_1 + v_2)$. The distribution of the sets of variables $(u_1^{(i)}, x_6^{(i)}, x_3^{(i)}, x_4^{(i)}, x_5^{(i)})$ is given by

$$f = \frac{1}{(2\pi)^{p/2}\sigma^p(v_1 + v_2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{p} (u_1^{(i)})^2 \right] \cdot \frac{1}{(2\pi)^{(n-1)p/2}|\mathbf{T}_1|^{p/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{p} (\mathbf{x}^{(i)} - \mathbf{x}_0^{(i)})(\sigma^2\mathbf{T}_1)^{-1}(\mathbf{x}^{(i)} - \mathbf{x}_0^{(i)})' \right], \tag{12}$$

where the $\mathbf{x}^{(i)}$ are the "reduced" vectors $(x_6^{(i)}, x_3^{(i)}, x_4^{(i)}, x_5^{(i)})$ and $\mathbf{T}_1$ is a tree matrix whose components correspond to the covariances among these variables. The columns and rows of $\mathbf{T}$ corresponding to populations 1 and 2 have been stricken, and a new column and row for "population" 6 has been added, so that

$$\mathbf{T}_1 = \left[\begin{array}{c|ccc} t_6' - t_0 & t_8 - t_0 & t_9 - t_0 & t_9 - t_0 \\ \hline t_8 - t_0 & & & \\ t_9 - t_0 & & \text{as before} & \\ t_9 - t_0 & & & \end{array}\right], \tag{13}$$

where $t'_6 = t_6 + v_1v_2/(v_1 + v_2)$. It will not matter where in the matrix the new row and column are inserted, provided that the $x_6^{(i)}$ are inserted in the corresponding position in the reduced vectors $\mathbf{x}^{(i)}$.

The result of these transformations is that when $x_1^{(i)}$ and $x_2^{(i)}$ are transformed to $u_1^{(i)}$ and $x_6^{(i)}$, the likelihood of the tree, given the transformed data, is the same as the product of the likelihoods of the two trees shown in figure 2. Populations 1 and 2 have been "pruned." They have been replaced by a one-population tree and by calculating a value for x_6 and a new time for point 6.

We can now repeat the process on the new tree. We could "prune" either populations 4 and 5 or populations 6 and 3. If we do the latter the transformations of phenotypes and times will be:

$$u_2^{(i)} = x_6^{(i)} - x_3^{(i)},$$

$$x_8^{(i)} = \left(\frac{v_3}{v_3 + v_6}\right) x_6^{(i)} + \left(\frac{v_6}{v_3 + v_6}\right) x_3^{(i)},$$

and (14)

$$t_8' = t_8 + \frac{v_3v_6}{(v_3 + v_6)},$$

where $v_3 = t_3 - t_8$ and $v_6 = t_6' - t_8$. This transformation leaves the values $u_1^{(i)}$, $x_4^{(i)}$, and $x_5^{(i)}$ unchanged. The likelihood of the original tree, given the doubly transformed set of phenotypes, is now the product of three independent distributions: one for the $u_1^{(i)}$, one for the $u_2^{(i)}$, and one for the populations $x_8^{(i)}$, $x_4^{(i)}$,

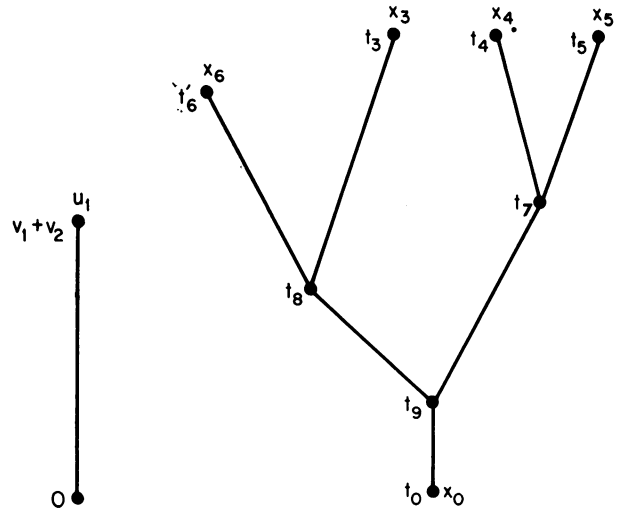

FIG. 2.—The tree after one step in the "pruning" process. The likelihood of the tree shown in fig. 1 is equal to the product of the likelihoods of the two trees shown here. We have $v_1 = x_1 - x_6$, $v_2 = x_2 - x_6$, $x_6 = (v_2x_1 + v_1x_2)/(v_1 + v_2)$, $t_6' = t_6 + v_1v_2/(v_1 + v_2)$, and $u_1 = x_1 - x_2$.

and $x_5^{(i)}$. Continuing this process, we ultimately have five sets of transformed values: $u_1^{(i)}$, $u_2^{(i)}$, $u_3^{(i)}$, $u_4^{(i)}$, and $x_9^{(i)}$. Their joint distribution is the product of five distributions:

$$
\begin{aligned}
f = {} & \frac{1}{(2\pi)^{p/2}\sigma^p(v_1+v_2)^{p/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{p}(u_1^{(i)})^2}{\sigma^2(v_1+v_2)}\right] \\
& \cdot \frac{1}{(2\pi)^{p/2}\sigma^p(v_6'+v_3)^{p/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{p}(u_2^{(i)})^2}{\sigma^2(v_6'+v_3)}\right] \\
& \cdot \frac{1}{(2\pi)^{p/2}\sigma^p(v_4+v_5)^{p/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{p}(u_3^{(i)})^2}{\sigma^2(v_4+v_5)}\right] \\
& \cdot \frac{1}{(2\pi)^{p/2}\sigma^p(v_8'+v_7')^{p/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{p}(u_4^{(i)})^2}{\sigma^2(v_8'+v_7')}\right] \\
& \cdot \frac{1}{(2\pi)^{p/2}\sigma^p(v_9')^{p/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{p}(x_9^{(i)}-x_0^{(i)})^2}{\sigma^2(v_9')}\right].
\end{aligned}
\tag{15}
$$

The likelihood of the original tree, given the transformed set of data, is the same as the product of the likelihoods of the five trees shown in figure 3. In that figure, as in equation (15), the v_i are primed if the corresponding t_i has been modified before it is used to calculate the v_i.

The reader who has followed the argument step by step may have detected some sleight of hand. We mean to calculate the likelihood of a tree based on a set of data, but equation (15) is the likelihood based on a transformed set of data. Will the result be affected? It can be shown that if an invertible linear transformation is applied to a set of data and if the original set of data has a continuous joint distribution, the probability density of the original values will be equal to the probability density of the transformed values times a constant, the Jacobian of the linear transformation. However, in this case, the Jacobians of the transformations given by equations (7) and (14) are unity, so that the probability density of a point in the original set of variables is the same as the probability density at the corresponding point in the transformed variables.

The last term of equation (15) raises an interesting set of problems. The $x_0^{(i)}$ must be estimated from the data. The maximum-likelihood estimate of $x_0^{(i)}$ is obviously $x_9^{(i)}$. Setting $x_0^{(i)} = x_9^{(i)}$, the last term depends on v_9'. This is $t_9' - t_0$. Since $t_9' > t_9$ as a result of the "pruning" process, we can make v_9' no smaller than $t_9' - t_9$. We are not confronted by a singularity—this term is always finite. But a problem does arise in the choice of t_0. If we make a maximum-likelihood estimate of t_0, then since it only enters into equation (15) as part of v_9', the maximum-likelihood estimate will be t_9. But t_0 is an arbitrary starting point, not a real prop-

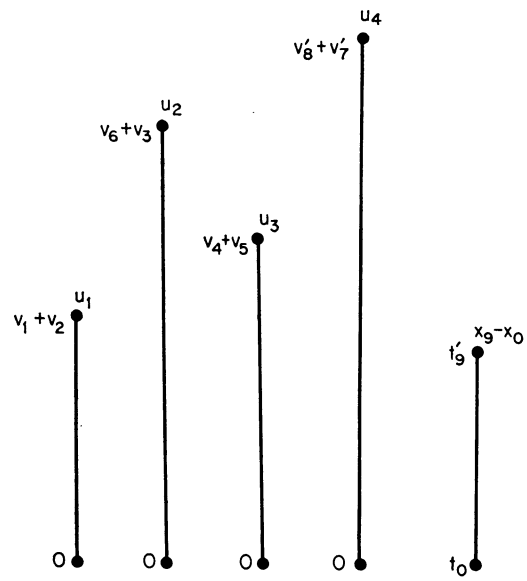

FIG. 3.—When the tree in fig. 1 has been completely "pruned," the likelihood of that tree, given the transformed set of data, is exactly the product of the likelihoods of these five single-population trees. See text for details.

erty of the evolutionary tree. Estimating t_0 from the data is very different from fixing it at some time. Note that

$$v_9' = t_9' - t_9 + t_9 - t_0. \tag{16}$$

If we take t_0 sufficiently far in the past, then the contribution of $(v_9')^{-p/2}$ to the likelihood given in equation (15) will not be much different for trees which differ in their values of t_9' or of t_9. In this case the last term of equation (15) will be effectively a constant, and it will not influence the choice of the tree. But if we estimate t_0, we get $t_0 = t_9$, so

$$v_9' = t_9' - t_9. \tag{17}$$

Then $(v_9')^{-p/2}$ will be greater the closer t_9' is to t_9. Now the last term of equation (15) has a substantial influence on the choice of a maximum-likelihood estimate.

In computer runs on actual data, the influence of using equation (17) usually forced the maximum-likelihood estimates to have a four-way fork at the base of the tree. When the last term of equation (15) was dropped, the maximum-likelihood estimates had a bifurcation at the bottom, and the four-way forks vanished. Since there is no particular reason to assume that true evolutionary trees preferentially have a multiway fork at their base, one would like some logical justification for using equation (16) and assuming $t_9 - t_0$ to be very large, in other words, for dropping the last term of equation (15). By dropping that term, we are in effect using only the $u_j^{(i)}$ to estimate the tree. These do not depend on the sizes of the $x_j^{(i)}$, but only on their differences. In effect, we are dropping the estimation of the $x_0^{(i)}$.

A compelling argument for doing this can be made in the case of a two-population tree. In that case, estimating the tree is the same as estimating the time of the first fork. If we use equation (17), it is easily shown that the maximum-likelihood estimate of this time is $\sum_i (x_1^{(i)} - x_2^{(i)})^2/4p$, while if we drop estimation of the $x_0^{(i)}$, the estimate of the time of the first fork is twice as large. The latter estimate is not only an unbiased estimate, it is also consistent. As we consider more and more characters the estimate converges on the true value. The estimate using equation (17) converges on half the true value. Therefore, in this case, the procedure of dropping the $x_0^{(i)}$ is clearly superior.

By considering only the $u_j^{(i)}$ and dropping estimation of the $x_0^{(i)}$ we are carrying out a "marginal-likelihood" procedure [7]. E. A. Thompson (personal communication) has shown that the procedure proposed here satisfies Kalbfleisch and Sprott's criteria for the use of marginal likelihoods.

The procedure outlined in this section applies also to trees containing trifurcations and branchings of higher order. To deal with them, it is only necessary to insert imaginary segments of zero length in such a way as to convert the multifurcations into bifurcations. This will not affect the likelihood of the tree. Nor will that likelihood become infinite as a result of the presence of segments of zero length. The v_i' in equation (15) are not, of course, the original lengths of the segments, but the lengths after t_i has been changed to t_i' by the "pruning" process. Since this

amounts to increasing v_i by half the harmonic mean of the final lengths of the two segments above segment i, v_i' will be positive. The only case in which a singularity can appear is the trivial one in which two branch tips have identical phenotypes in all characters.

THE SUFFICIENCY OF DISTANCES

The quantities $\Sigma(u_k^{(i)})^2$ in equation (15) bear a suspicious resemblance to distances. This raises the question of whether we can compute the likelihood (15) directly from the pairwise distances between the populations. The reader must already know the answer, since I obviously would not have raised the question if it were not possible to confine our attention to distances.

If we plot each population in a space defined by the values of its phenotypes, so that the coordinates of population j are $(x_j^{(1)}, \ldots, x_j^{(p)})$, the distance between populations 1 and 2 is indeed given by

$$D_{12}^2 = \sum_{i=1}^{p} (x_1^{(i)} - x_2^{(i)})^2 = \sum_{i=1}^{p} (u_1^{(i)})^2. \tag{18}$$

Then if we know D_{12}^2, we can substitute it directly into the first term of equation (15). An analogous principle holds for any pair of tip populations descended from the same immediate ancestor. Thus, we can also substitute D_{45}^2 directly into the third term of equation (15). However, the corresponding quantities in the second and fourth terms of equation (15) are not equal to any of the D_{ij}^2. The quantity in the second term is

$$\sum_{i=1}^{p} (u_2^{(i)})^2 = \sum_{i=1}^{p} (x_6^{(i)} - x_3^{(i)})^2. \tag{19}$$

The $x_6^{(i)}$ are not original data points, but are averages of the $x_1^{(i)}$ and $x_2^{(i)}$, namely,

$$x_6^{(i)} = \left(\frac{v_2}{v_1 + v_2}\right) x_1^{(i)} + \left(\frac{v_1}{v_1 + v_2}\right) x_2^{(i)}. \tag{20}$$

We can show by straightforward but tedious algebra that if

$$D_{63}^2 = \sum_{i=1}^{p} (x_6^{(i)} - x_3^{(i)})^2,$$

then

$$D_{63}^2 = \left(\frac{v_2}{v_1 + v_2}\right) D_{13}^2 + \left(\frac{v_1}{v_1 + v_2}\right) D_{23}^2 - \left(\frac{v_1}{v_1 + v_2}\right)\left(\frac{v_2}{v_1 + v_2}\right) D_{12}^2. \tag{21}$$

More generally, if we form a new population, k, by "pruning" the segments i and j immediately above fork k, the distance from the new "population" to some other population, m, is given by

$$D_{km}^2 = \left(\frac{v_j}{v_i + v_j}\right) D_{im}^2 + \left(\frac{v_i}{v_i + v_j}\right) D_{jm}^2 - \left(\frac{v_i}{v_i + v_j}\right)\left(\frac{v_j}{v_i + v_j}\right) D_{ij}^2. \tag{22}$$

Every time we "prune" the tree we can use equation (22) to calculate the distance from the new tip population to all the other tip populations. If we drop the last term of equation (15), that equation becomes

$$f = \frac{1}{(2\pi)^{4p/2}\sigma^{4p}[(v_1 + v_2)(v_6' + v_3)(v_4 + v_5)(v_8' + v_7')]^{p/2}} \cdot \exp\left[-\frac{1}{2\sigma^2}\left(\frac{D_{12}^2}{v_1 + v_2} + \frac{D_{63}^2}{v_6' + v_3} + \frac{D_{45}^2}{v_4 + v_5} + \frac{D_{87}^2}{v_8' + v_7'}\right)\right]. \tag{23}$$

The likelihood of this tree depends on the original data only through the quantities D_{12}^2, D_{63}^2, D_{45}^2, and D_{87}^2. The quantities D_{12}^2 and D_{45}^2 are distances between original tip populations, and both D_{63}^2 and D_{87}^2 can be calculated from the original pairwise distances D_{ij}^2. In fact, all of the D_{ij}^2 enter into equation (23) in one way or another. Provided that dropping the last term in equation (15) is valid, the set of pairwise distances D_{ij}^2 are sufficient statistics for determining the maximum-likelihood evolutionary tree.

CORRELATED CHARACTERS

The above distances were calculated from characters which were assumed to undergo independent Brownian motion with equal variance, σ^2. We often encounter characters with functional or genetic correlations. We usually have no direct information about the covariances of the evolutionary changes in the characters, but frequently samples of individuals from each population are available, enabling estimation of the within-population covariances of the characters. The generalized distances (or "Mahalanobis distances") can then be calculated between all pairs of populations. There are two conditions which must be satisfied (or at least closely approximated) if generalized distances are to be used as the distances between populations in the above analysis. First, the mean population phenotype for each character must be a linear combination of a series of underlying variables, each of which changes through time according to an independent Brownian motion process. Second, the covariances of the evolutionary changes in the observed characters must be proportional to the within-population covariances of these characters. When we do not know the constant of proportionality between the two sets of covariances, this will be equivalent to not knowing σ^2 in the underlying Brownian motion. Therefore when we obtain an estimate of the evolutionary tree, the times of branching can be estimated only in units of $1/\sigma^2$, and these cannot be converted to years from the present unless σ^2 is known.

In calculating the generalized distances, we are in effect finding a linear transformation which makes the characters independent and of unit variance each. This transformation which carries us from the correlated observed characters $\mathbf{y}^{(i)}$ to the independent characters $\mathbf{x}^{(i)}$ does affect the likelihood. Equations such as (23) calculate the probability density of the transformed variables, which equals the probability density of the original variables divided by the Jacobian of the transformation. In this case, the Jacobian is not necessarily unity. But the transformation is independent of the particular evolutionary tree. Since all statistical conclusions are drawn from the ratios of the likelihoods of different evolutionary trees, the Jacobians will cancel out and the result will therefore be unaffected by the transformation. In practice, of course, we never actually obtain the transformation: it is implicit in the calculation of the D_{ij}^2.

A more systematic statement of the algorithm for calculating the likelihood is given in the Appendix. Combining this algorithm with routines for making small changes in the shape of an evolutionary tree, it is possible to construct computer programs which search for a local maximum in the likelihood surface and thus make a maximum-likelihood estimate of the evolutionary tree. This sort of procedure does not guarantee that we will find the true maximum-likelihood estimate of the tree topology. It does not even guarantee that if our final estimate has a certain topology, the times of branching will be the best which could accompany that particular topology. It will have the same limitation as any "hill-climbing" maximization algorithm: it will climb the nearest hill rather than the highest one. The more rearrangements of the tree that are tried and the more runs that are made with different initial trial trees, the more likely it is that the program will reach the highest point on the likelihood surface. This paper presents only an algorithm for calculating the likelihood. Since I claim no special validity for the routines I use to make small changes in an evolutionary tree, readers are left to their own devices in constructing computer programs.

QUANTITATIVE CHARACTERS

We now consider whether polygenic quantitative characters satisfy the assumptions made above, so that generalized distances based on these characters can be used in the estimation of evolutionary trees. First, consider an additively determined trait with no dominance. The character is determined by m loci, each having two alleles. The contribution of the ith locus to the individual's phenotype is $2a_i$, a_i, or 0, depending on whether the genotype at that locus is $A^{(i)}A^{(i)}$, $A^{(i)}a^{(i)}$, or $a^{(i)}a^{(i)}$. The population mean of the character is then

$$\mu = \sum_{i=1}^{m} 2p_i a_i, \tag{24}$$

where p_i is the gene frequency of $A^{(i)}$. The population mean is thus a linear combination of m random variables, the p_i.

The change in the mean in one generation is

$$\Delta\mu = 2\sum_{i=1}^{m}(\Delta p_i)a_i. \tag{25}$$

Assume that all changes are the result of random genetic drift, with effective population size N_e. Then

$$E(\Delta\mu) = 2\sum_{i=1}^{m}E(\Delta p_i)a_i = 0; \tag{26}$$

$$\text{var}(\Delta\mu) = 4\sum_{i=1}^{m}\text{var}(\Delta p_i)a_i^2 = 4\sum_{i=1}^{m}\frac{p_i(1-p_i)}{2N_e}a_i^2. \tag{27}$$

We need to check whether the variance in evolutionary change (27) will be proportional to the within-population variance. If there is no environmental contribution to the character, the within-population variance is

$$\text{var}\,(P) = \sum_{i=1}^{m}2p_i(1-p_i)a_i^2, \tag{28}$$

so that the variance of evolutionary change is exactly $1/N_e$ times the within-population variance. To be able to use generalized distances, we must also verify that within-population variances are the same. This will be true if (1) the gene frequencies in different populations, p_i, are not far apart; or (2) the character is controlled by so many loci that if a population has $p_i(1-p_i)$ at one locus unusually large, it will have other loci with $p_i(1-p_i)$ unusually small, so that equation (28) will be approximately constant.

If the effective numbers, N_e, of the populations are unequal, their variances of evolutionary change will be unequal. Nevertheless, it is still possible to carry out maximum-likelihood estimation. If one population has half the effective population number of another, its variance of gene frequency change will be twice as large. We can therefore interpret our "time" scale as actually measuring the total amount of evolution, proportional to the sum of $1/N_e$ over generations. Since some tip populations will have accumulated a larger amount of genetic drift than others, we must also allow the "times" of the tip populations to be unequal. We can do this by fixing one tip population's "time" arbitrarily, and allowing all the other tip population "times" to be estimated. It can be shown that when we do this, we will not be estimating a rooted evolutionary tree, but instead an unrooted tree (which may also be thought of as a branching network containing no loops).

Sometimes the same genes will contribute to several characters. This will not create any difficulty. If we treat the gene dose at each locus as a separate character, it can assume the values 0, 1, and 2. Each gene dosage will have the property that its evolutionary variance is $1/N_e$ times its within-population variance. Furthermore, *any* linear combination of the gene doses will have this property. If we find a set of linear combinations of the phenotypes which are independent, as we always can, these transformations will preserve the property that the variance of evolutionary change is proportional to within-population variance.

It was assumed that the characters had no environmental contribution. If they are affected by environmental factors, we can write $V_T = V_A + V_E$, where V_T is the within-population total variance, V_A is the quantity calculated in equation (28), and V_E is environmental variance. In calculating D^2, we want to use the quantity in equation (28), which is V_A. Neither equation (27) nor the covariances between characters will be affected by the presence of environmental variance (in simple cases). We can multiply observed within-population variance by h^2 $(= V_A/V_T)$, the heritability of the character, to obtain the additive variance V_A. After V_T has been changed to V_A for each character by using the appropriate h^2, we can proceed to calculate D^2 using this altered within-population covariance matrix. If the environmental components of different characters are correlated, the correction is somewhat more complicated. Environmental differences between populations can also mimic genetic differences. The lack of information about possible environmental effects within and between populations will be a serious source of error with many types of quantitative character data. Persons analyzing such data should proceed with extreme caution.

So far, all genetic variance has been additive. If we allow the presence of dominance variance, the situation is more complicated. The mean phenotype is now a quadratic function of the individual gene frequencies. This means that genetic drift in the frequencies will cause a change in the mean phenotype, a change whose expectation is not zero. One might imagine that the presence of this "inbreeding depression" would invalidate a Brownian motion model. However, if the gene frequencies in the individual populations are nearly the same, the amount of phenotypic change due to inbreeding depression will be approximately the same in all populations. In fact, it can be shown that for large effective population number, the differences between populations will behave the same way as do the differences between populations in the case of Brownian motion. The variance of evolutionary change is equal to $1/N_e$ times the additive genetic variance of the character. Therefore the method of correcting within-population variances remains the same: multiply them by the heritability of the character. These results can be extended to cases with multiple alleles, but this will not be done in this paper.

In these models, the random component of phenotypic change has been due to random genetic drift. At first sight, it appears that we can make much the same analysis if the randomness is due to variation in selection coefficients. However, in this case there is no natural relationship between within-population covariances and the covariances of evolutionary change. Characters which are independent within a population may have selection pressures which are highly correlated, and vice versa. There are also mathematical differences in the form of the equations for variances of gene frequencies. These cast further doubt on the tractability of models of "selective drift."

GENE FREQUENCY DATA

Edwards and Cavalli-Sforza [1] introduced measures of distance calculated from the frequencies of alleles at a number of polymorphic loci. For a locus with two

alleles, the allele dose per gene, P, can be considered as an additive phenotype with $a_i = \frac{1}{2}$ and no dominance or environmental variance. The mean gene dose is the gene frequency of the population. The within-population variance of gene dose is

$$\text{var}(P) = \frac{1}{2} p_i(1 - p_i), \tag{29}$$

and the variance of the evolutionary change in one generation is

$$\text{var}(\Delta P) = \text{var}(\Delta p_i) = \frac{p_i(1 - p_i)}{2N_e}, \tag{30}$$

so that gene frequencies under random genetic drift satisfy the condition that the variance of the evolutionary change is $1/N_e$ times the within-population variance. If the p_i are sufficiently close to each other that $p_i(1 - p_i)$ is nearly the same in all populations, the requirements for approximating this process by Brownian motion will be met.

A measure equivalent to D^2 in this case is

$$\frac{(p_1 - p_2)^2}{\frac{1}{2}\bar{p}(1 - \bar{p})} = 2\left[\frac{(p_1 - p_2)^2}{\bar{p}} + \frac{(p_1 - p_2)^2}{1 - \bar{p}}\right], \tag{31}$$

where p_1 and p_2 are the frequencies of one allele in the two populations and $\bar{p}$ is the arithmetic mean of p_1 and p_2. Kurczynski [8] has shown that for equal sample sizes in the case of multiple alleles, an analogue to equation (31) is

$$\sum_{j=1}^{n} \frac{2(p_{j1} - p_{j2})^2}{\frac{1}{2}(p_{j1} + p_{j2})}, \tag{32}$$

where p_{j1} and p_{j2} are the frequencies of allele j in populations 1 and 2. Summation is over all alleles at the locus. I have multiplied Kurczynski's actual formula by 2 to make it directly comparable with equation (31). Kurczynski's original formula would be applicable in a haploid population. In a population of diploids, if D^2 is to have the same approximate meaning for a quantitative character as for a gene frequency, we must use equation (31) or (32) or approximately equivalent formulas.

Other measures of genetic distance have been proposed [1, 9–11]. All of these distance measures have the property that they are asymptotically equivalent to each other when the p_{j1} are close to p_{j2}, and their properties differ when this does not hold. This area in which their general properties are similar is also the area in which the random genetic drift process most nearly satisfies the assumptions of a Brownian motion model. From a maximum-likelihood standpoint, that measure of genetic distance is best which comes closest to yielding the correct solution for the maximum-likelihood evolutionary tree. At present, it is not known which measure of genetic distance is best. From the point of view of this paper, it does not matter much which is used, provided that one knows how to combine measures at different

loci and how to combine distances for gene frequency data with values of D^2 for quantitative characters.

In combining gene frequency distances with quantitative character distances, a useful general principle is that the expected value of D^2 for a single quantitative character must be the same as the expected value for a two-allele locus (since a two-allele locus should give the same information if we consider the dosage of one allele to be a quantitative character). Thus, both are expected to contribute equal amounts of information. More generally, the expected value of D^2 after one generation of divergence should be $2k/N_e$, where k is the number of degrees of freedom in the character being measured. For quantitative characters, k is the number of characters (excluding cases of complete dependence of some characters on others), and for gene frequencies it is one less than the number of alleles.

For gene frequencies in diploids, the measures G_s^2, B^2, and G_c^2 given by Balakrishnan and Sanghvi [9], and the measure D_k^2 given by Kurczynski [8], all have expectations k/N_e for two populations which have been separated for one generation. Thus, they can be used in place of D^2 once they have been multiplied by 2. The factor of 2 can be justified on intuitive grounds. All of these distance measures use as their equivalents of within-population allele covariances the covariances of allele doses for a single gene. To be comparable with quantitative characters one would have to use the covariances of average allele doses in a diploid genome. This differs by a factor of 2.

The distance measure E^2 proposed by Edwards [11, 12] cannot be used directly in place of D^2. Since

$$E^2 = \frac{8[1 - \sum_i (p_{i1}p_{i2})^{1/2}]}{[1 + \sum_i (\bar{p}_i/n)^{1/2}][1 + \sum_i (\bar{p}_i/n)^{1/2}]}, \tag{33}$$

the expected value of E^2 after one generation of drift in both populations is approximately

$$E(E^2) = \frac{(n-1)}{N_e}/[1 + 2\sum_i (\bar{p}_i/n)^{1/2} + 1/n], \tag{34}$$

ignoring terms of order N_e^{-2}. Thus the expected divergence is dependent on the gene frequencies. The dependence arises from the presence of the denominator in equation (33). If we use only the numerator of equation (33) as our measure of distance, it has expectation $(n-1)/N_e = k/N_e$, and in fact is an exact multiple of the measure proposed by Edwards and Cavalli-Sforza [1]. Since the numerator must be multiplied by 2 to be comparable with D^2, their distance measure should be used in the form

$$S = 16\left[1 - \sum_{i=1}^{n} (p_{i1}p_{i2})^{1/2}\right]. \tag{35}$$

The distance D of Nei [10] can also be shown to have expected divergence dependent on the gene frequencies. For small differences between p_{i1} and p_{i2}, Edwards

and Cavalli-Sforza's S is asymptotically the same as G_s^2, B^2, G_c^2, and D_k^2. Which of these measures performs best in maximum-likelihood calculations when inserted in place of D^2 has not been clearly determined.

STANDARD ERRORS FOR THE ESTIMATES

As in any maximum-likelihood procedure, we can calculate asymptotic variances and covariances of our estimates by taking $-\mathbf{A}^{-1}$, where $\mathbf{A}$ is the matrix of second partial derivatives of the log-likelihood surface at the maximum-likelihood values. It is possible to obtain separate expressions for each of these second derivatives using matrix derivatives. The evaluation of $-\mathbf{A}^{-1}$ will then involve about as much calculation as obtaining $m^2/2$ likelihoods, where there are m times being estimated. A more straightforward approach involving the same amount of calculation would be to calculate second-order differences in the log likelihood for small displacements around the estimates, use these as approximations to the elements of $\mathbf{A}$, and then obtain $-\mathbf{A}^{-1}$ by direct inversion, as before. This also involves calculation of approximately $m^2/2$ likelihoods.

Both of these procedures may involve substantial computation compared with the computation necessary to find the maximum-likelihood tree in the first place. A cruder but simpler technique is to calculate the second-order difference of the log likelihood with respect to variation in each of the m parameters separately. Taking the inverse of each of these, we get a quantity which is an underestimate of the true asymptotic variance for that parameter. These quantities will be useful (1) to give a rough idea as to which parts of the tree are best known, and (2) to provide lower bounds for the true variances. These standard errors relate only to the time of each node. The implicit assumption is that a sufficiently large amount of data has been gathered so that the topological form of the tree is known and that only the exact fork times remain to be determined. This will almost never be true in practice. Therefore, large standard errors of node times must be taken as indicators of uncertainty about the topology of the tree.

It is easy to overinterpret the results of this sort of analysis. Fork-time variances and covariances should be taken, as Mark Twain said, "with a ton of salt." The validity of such variances and covariances is dependent on the correctness of the underlying model of evolution. Surely we are justified in being highly skeptical of these models. When the data are based on quantitative characters, it will be necessary to make size corrections, since size may change more readily in evolution than shape. We must also be persuaded that the measurable characters are not subject to strong natural selection, a most dubious assumption.

When gene frequency data are used, we are usually dealing, not with separate species, but with populations which are only partially reproductively isolated. A model of evolution by branching and complete reproductive isolation will be inappropriate. Morton et al. [13] have been particularly critical of the use of branching models for the evolution of human populations. Real human populations have neither maintained the complete isolation implicit in branching models nor the steady rates of exchange of migrants implicit in models of "isolation by distance."

A more realistic model of human populations might prove to be computationally intractable. Until such a model can be developed, branching models and isolation-by-distance models will be most useful when both are fitted to the same data.

SUMMARY

Edwards and Cavalli-Sforza proposed the estimation of evolutionary trees by maximum likelihood for a Brownian motion model of evolutionary change. They were prevented from calculating such estimates by singularities in their likelihood function. It is shown that if one drops the estimation of the phenotypes of the fork populations and estimates only fork times, there are no singularities in the resulting likelihood surface, and ordinary maximum-likelihood estimation is possible. Estimation of the initial phenotype and time of the initial population is dropped using a procedure equivalent to "marginal likelihood." It is then shown that the generalized distances between all pairs of populations are sufficient statistics for the estimation of the maximum-likelihood tree. A simplified computational procedure is derived to calculate the likelihood of an evolutionary tree. It is shown that quantitative characters and gene frequencies approximately satisfy the assumptions of the model of evolution by Brownian motion if evolution is by random genetic drift. Generalized distances calculated from these two types of data can be combined, and guidelines for doing this are given. The interpretation of standard errors obtained from the maximum-likelihood procedure is discussed.

ACKNOWLEDGMENTS

I am deeply indebted to A. W. F. Edwards. My initial interest in this problem originated from a conversation with him, and he has made numerous helpful suggestions and criticisms throughout the work. I wish to express my gratitude to my thesis adviser, R. C. Lewontin, for displaying a degree of tolerance unusual even for him. I wish to thank Harvey Motulsky for writing the first computer programs to carry out these procedures and for his well-founded suspicion of four-way branchings. I also wish to thank L. L. Cavalli-Sforza, E. A. Thompson, and N. E. Morton for their constructive criticisms; K. K. Kidd for his encouragement; and M. Kimura, T. Maruyama, T. Ohta, and T. Yamazaki of the National Institute of Genetics, Japan, for their hospitality during the writing of this paper.

APPENDIX
PROCEDURE FOR CALCULATING LIKELIHOOD OF A TREE FROM GENERALIZED DISTANCES

We are given an $n \times n$ array, D_{ij}^2, of the squared generalized distances between tip populations; the number of characters, p; the times, t_i, of the tip populations and of the forks (which may be either bifurcations or multiway branchings); a function, anc(i), which gives the number of the fork immediately ancestral to each fork or tip i, and an array of indicator numbers, or some other means of indicating which tip populations are still on the evolutionary tree (i.e., have not yet been "pruned"). It is assumed that the times, t_i, have already been multiplied by σ^2, so that one generation has become σ^2 units of "time." Then the following procedure will calculate the likelihood of the tree:

1. Let $S = 0$ and $T = 1$.

2. Find two tip populations on the tree, say i and j, which have the same immediate ancestor, k, so that $k = \text{anc}(i) = \text{anc}(j)$.

3. Let $v_1 = t_i - t_k$, $v_2 = t_j - t_k$, and $f = v_1/(v_1 + v_2)$.

4. Change S and T: $S \leftarrow S + D_{ij}^2/(v_1 + v_2)$, $T \leftarrow T\,(v_1 + v_2)$, where "$\leftarrow$" means "is replaced by."

5. Recalculate the squared distances, D_{im}^2, between population i and every other population m still on the tree (except population j): $D_{im}^2 \leftarrow (1 - f)D_{im}^2 + fD_{jm}^2 - f(1 - f)D_{ij}^2$. Be sure to keep the new values of D_{mi}^2 and D_{im}^2 equal.

6. Remove tip j from the tree.

7. Change the time of tip i to $t_i' = t_k + v_1v_2/(v_1 + v_2)$.

8. If there are no further tips or forks which have k as their ancestor, other than i and j (i.e., if k was a bifurcation), remove fork k from the tree and change anc(i) so that it is now equal to the previous value of anc(k).

9. Go back to step 2 unless there is now only one tip remaining on the tree. If so, then we can now calculate the likelihood as $L = T^{-p/2}\, e^{-S/2}$. A factor of $(2\pi)^{-(n-1)p/2}$ has been omitted from this expression. If log likelihood is preferred, it is $\log_e L = -(p/2)\log_e T - S/2$.

Since T may be a product of small quantities, it might be preferable to replace it by W, where initially $W = 0$, and in step 4: $W \leftarrow W + \log_e(v_1 + v_2)$. Then $\log_e L = -(p/2)W - S/2$.

Note that the procedure given here differs from the "pruning" process described in the text of this paper in that when populations i and j are removed by "pruning," the synthetic population which replaces them is called i rather than k. This will not, of course, affect the result.

REFERENCES

1. Edwards AWF, Cavalli-Sforza LL: Reconstruction of evolutionary trees, in *Phenetic and Phylogenetic Classification,* edited by Heywood VH, McNeill J, London, Systematics Association Publication no. 6, 1964, pp 67–76
2. Cavalli-Sforza LL, Edwards AWF: Analysis of human evolution, in *Genetics Today, Proceedings 11th International Congress of Genetics,* vol 3, edited by Geerts SJ, Oxford, Pergamon, 1965, pp 923–933
3. Cavalli-Sforza LL, Edwards AWF: Phylogenetic analysis: models and estimation procedures. *Evolution* 21:550–570, 1967; *Am J Hum Genet* 19:233–257, 1967
4. Edwards AWF: Estimation of the branch-points of a branching-diffusion process. *J R Statist Soc B* 32:155–174, 1970
5. Kidd KK, Sgaramella-Zonta LA: Phylogenetic analysis: concepts and methods. *Am J Hum Genet* 23:235–252, 1971
6. Malyutov MB, Passekov VP, Rychkov YG: On the reconstruction of evolutionary trees of human populations resulting from random genetic drift, in *The Assessment of Population Affinities in Man,* edited by Weiner JS, Huizinga J, Oxford, Clarendon, 1972, pp 48–71
7. Kalbfleisch JD, Sprott DA: Application of likelihood methods to models involving large numbers of parameters. *J R Statist Soc B* 32:175–208, 1970
8. Kurczynski TW: Generalized distance and discrete variables. *Biometrics* 26:525–534, 1971
9. Balakrishnan V, Sanghvi LD: Distance between populations on the basis of attribute data. *Biometrics* 24:859–865, 1968
10. Nei M: Genetic distance between populations. *Am Naturalist* 106:283–292, 1972

11. EDWARDS AWF: Distances between populations on the basis of gene frequencies. *Biometrics* 27:873–881, 1971
12. EDWARDS AWF, CAVALLI-SFORZA LL: Affinity as revealed by differences in gene frequencies, in *The Assessment of Population Affinities in Man,* edited by WEINER JS, HUIZINGA J, Oxford, Clarendon, 1972, pp 37–47
13. MORTON NE, YEE S, HARRIS DE, LEW R: Bioassay of kinship. *Theor Pop Biol* 2:507–524, 1971